Utlity-based Dueling Bandits as a Partial Monitoring Game

Pratik Gajane pratik.gajane@inria.fr
INRIA SequeL team, Lille 59650, France
Orange labs, Lannion 22300, France

Tanguy Urvoy tanguy.urvoy@orange.com
Orange labs, Lannion 22300, France

Editor:

Abstract

Partial monitoring is a generic framework for sequential decision-making with incomplete feedback. It encompasses a wide class of problems such as dueling bandits, learning with expect advice, dynamic pricing, dark pools, and label efficient prediction. We study the utility-based dueling bandit problem as an instance of partial monitoring problem and prove that it fits the time-regret partial monitoring hierarchy as an *easy* – i.e. $\tilde{\Theta}\left(\sqrt{T}\right)$ – instance. We survey some partial monitoring algorithms and see how they could be used to solve dueling bandits efficiently.

Keywords: Online learning, Dueling Bandits, Partial Monitoring, Partial Feedback, Multiarmed Bandits

1. Introduction

Partial Monitoring (PM) provides a generic mathematical model for sequential decision-making with incomplete feedback. It is a recent paradigm in the reinforcement learning community. Similarly the multi-armed bandit problem is a classical mathematical model for the *exploration/exploitation* dilemma inherent in reinforcement learning (see Bubeck and Cesa-Bianchi, 2012). The *K-armed dueling bandit problem* (Yue and Joachims, 2009) is a variation of the multi-armed bandit problem where two arms are selected at each round with a relative feedback.

Several generic partial monitoring algorithms have been proposed for both stochastic and adversarial settings (see Bartók et al., 2014, for details). With the exception of globalexp3 Bartók (2013) which tries to capture the structure of the games more finely, these algorithms only focus on the time bound and perform inefficiently in term of the number of actions. As we show in section 5, for a dueling bandit problem, the number of actions is quadratic in the number of arms K and these algorithms, including globalexp3, provide at best a $\tilde{\mathcal{O}}\left(K\sqrt{T}\right)$ regret guarantee whereas a dedicated algorithm like rex3 (Gajane et al., 2015) can provide a $\tilde{\mathcal{O}}\left(\sqrt{KT}\right)$ guarantee[1]. Studying partial monitoring algorithms from the perspective of dueling bandits is hence an interesting and challenging problem which could help us improve the ability of PM algorithms to capture the structure of sequential decision problems in a better way.

1. The $\tilde{\mathcal{O}}\left(\cdot\right)$ notation hides logarithmic factors.

In this preliminary work, we investigate how a utility-based dueling bandits problem can be modeled as an instance of a partial monitoring game. Our main contribution is that, we prove, using the PM formalism, that it is an *easy PM instance* according to the hierarchy defined in Bartók et al. (2014). Furthermore, we take a brief look at the existing partial monitoring algorithms and examine how they could be used to solve dueling bandits problems efficiently.

1.1 Dueling bandits

The K-armed dueling bandit problem is a variation of the classical multi-armed bandit-problem introduced by Yue and Joachims (2009) to formalize the exploration/exploitation dilemma in learning from preference feedback. In its utility-based formulation, at each time period, the environment sets a bounded value for each of the K arms and simultaneously the learner selects two arms. The learner only sees the outcome of the *duel* between the selected arms (i.e. the feedback indicates which of the selected arms has better value) and receives the average of the gains of the selected arms. The goal of the learner is to maximize her *cumulative gain*.

Relative feedback is naturally suited to many practical applications because users are more obliging to provide a relative preference feedback rather than an absolute feedback e.g. compared to “I rate Tennis at 32/50 and Football at 48/50” (absolute feedback) , it’s easier for users to say “I like Football more than Tennis” (relative feedback). Information Retrieval systems with *implicit feedback* are another important application of the dueling bandits (see Radlinski and Joachims, 2007). The major difficulty of the dueling bandit problem is that the learner cannot directly observe the loss (or gain) of the selected actions. To capture this aspect of the problem, it can be modeled as an instance of the *partial monitoring problem* as defined by Piccolboni and Schindelhauer (2001).

1.2 Partial monitoring games

A partial monitoring game is defined by a tuple $(\boldsymbol{N}, \boldsymbol{M}, \boldsymbol{\Sigma}, \mathcal{L}, \mathcal{H})$ [2] where $\boldsymbol{N}$, $\boldsymbol{M}$, and $\boldsymbol{\Sigma}$ are the *action* set, the *outcome* set, and the *feedback alphabet* respectively. To each action $I \in \boldsymbol{N}$ and outcome $J \in \boldsymbol{M}$, the *loss function* $\mathcal{L}$ associates a real-valued loss $\mathcal{L}(I, J)$ and the *feedback function* $\mathcal{H}$ associates a feedback symbol $\mathcal{H}(I, J) \in \boldsymbol{\Sigma}$.

In every round, the opponent and the learner simultaneously choose an outcome J_t from $\boldsymbol{M}$ and an action I_t from $\boldsymbol{N}$, respectively. The learner then suffers the loss $\mathcal{L}(I_t, J_t)$ and receives the feedback $\mathcal{H}(I_t, J_t)$. Only the feedback is revealed to the learner, the outcome and the loss remain hidden. In some problems, gain $\mathcal{G}$ is considered instead of loss. The loss function $\mathcal{L}$ and the feedback function $\mathcal{H}$ are known to the learner. When both $\boldsymbol{N}$ and $\boldsymbol{M}$ are finite, the loss function and the feedback function can be encoded by matrices, namely loss matrix and feedback matrix each of size $|\boldsymbol{N}| \times |\boldsymbol{M}|$. The aim of the learner is to control the expected cumulative regret against the best single-action (or pure) strategy at time T:

$$R_T = \max_i \sum_{t=1}^{T} \mathcal{L}(I_t, J_t) - \mathcal{L}(i, J_t)$$

2. Uppercase boldface letters are used to denote sets

Various interesting problems can be modeled as partial monitoring games, such as learning with expect advice (Littlestone and Warmuth (1994)), the multi-armed bandit problem (Auer et al. (2002)), dynamic pricing (Kleinberg and Leighton (2003)), the dark pool problem (Agarwal et al. (2010)), label efficient prediction (Cesa-bianchi et al. (2005)), and linear and convex optimization with full or bandit feedback (Zinkevich (2003), Abernethy et al. (2008), Flaxman et al. (2004)). We shall briefly explain a couple of examples:

The dynamic pricing problem: A seller has a product to sell and the customers wish to buy it. At each time period, the customer secretly decides on a maximum amount she is willing to pay and the seller sets a selling price. If the selling price is below the maximum amount the buyer is willing to pay, she buys the product and the seller's gain is the selling price she fixed. If the selling price is too expensive, her gain is zero. The feedback is partial because the seller only recieves a binary information stating whether the customer has bought the product or not. A PM formulation of this problem is provided below:

$$x \in \boldsymbol{N} \subseteq \mathbb{R}, \quad y \in \boldsymbol{M} \subseteq \mathbb{R}, \quad \boldsymbol{\Sigma} = \{\text{“sold”}, \text{“not sold”}\}$$

$$\mathcal{G}(x,y) = \begin{cases} 0, & \text{if } x > y, \\ x, & \text{if } x \leq y, \end{cases} \qquad \mathcal{H}(x,y) = \begin{cases} \text{“not sold”}, & \text{if } x > y, \\ \text{“sold”}, & \text{if } x \leq y, \end{cases}$$

The multi-armed bandit problem: At each time period, the learner pulls one of the K arms and receives it's corresponding gain which is bounded in $[0,1]$. The learner sees only her gain and not the gain of other arms. The learner's goal is to win almost as much as the optimal arm. A partial monitoring formulation of this problem is provided with a set of K arms/actions $i \in \boldsymbol{N} = \{1,\dots,K\}$, an alphabet $\boldsymbol{\Sigma} = [0,1]$, and a set of environment outcomes which are vectors[3] $\boldsymbol{m} \in \boldsymbol{M} = [0,1]^K$. The entry with index i ($\boldsymbol{m}_i$) denotes the instantaneous gain of the i^{th} arm. Assuming binary gains, $\boldsymbol{M}$ is finite and of size 2^K.

$$\mathcal{G}(i,\boldsymbol{m}) = \boldsymbol{m}_i \qquad \mathcal{H}(i,\boldsymbol{m}) = \boldsymbol{m}_i$$

2. Dueling bandits as a Partial Monitoring game

The utility-based dueling bandits model is similar to multi-armed bandits but the action sets differ. An action consists here of selecting a pair (i,j) of arms. However, symmetric actions like (i,j) and (j,i) lead to the same gains and provide equally informative feedback. Hence the action set for the learner can be restricted to $\boldsymbol{N} = \{(i,j) : 1 \leq i,j \leq K, i \leq j\}$. When the environment selects an outcome $\boldsymbol{m} \in \boldsymbol{M}$ and the learner selects a duel/action $(i,j) \in \boldsymbol{N}$, the instantaneous gain $\mathcal{G}((i,j),\boldsymbol{m})$ and feedback $\mathcal{H}((i,j),\boldsymbol{m})$ are as follows:

$$\mathcal{G}((i,j),\boldsymbol{m}) = \frac{\boldsymbol{m}_i + \boldsymbol{m}_j}{2} \qquad \mathcal{H}((i,j),\boldsymbol{m}) = \begin{cases} \square & \text{if } \boldsymbol{m}_i < \boldsymbol{m}_j \quad \textit{(loss)} \\ \diamond & \text{if } \boldsymbol{m}_i = \boldsymbol{m}_j \quad \textit{(tie)} \\ \blacksquare & \text{if } \boldsymbol{m}_i > \boldsymbol{m}_j \quad \textit{(win)} \end{cases}$$

To illustrate this formalism, we encode a 4-armed binary-gain dueling bandit problem as a PM problem in Figure 1. The first element of every column is of the form $\boldsymbol{m}_1\boldsymbol{m}_2\boldsymbol{m}_3\boldsymbol{m}_4$ where $\boldsymbol{m}_i$ is the gain for i^{th} arm. The first element of every row is of the form d_1d_2 where d_1 is the first arm being picked and d_2 being the second.

3. Lowercase boldface letters are used to denote vectors

$\mathcal{G}$ =

	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
11	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
12	0	0	0	0	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1	1	1	1
13	0	0	1/2	1/2	0	0	1/2	1/2	1/2	1/2	1	1	1/2	1/2	1	1
14	0	1/2	0	1/2	0	1/2	0	1/2	1/2	1	1/2	1	1/2	1	1/2	1
22	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
23	0	0	1/2	1/2	1/2	1/2	1	1	0	0	1/2	1/2	1/2	1/2	1	1
24	0	1/2	0	1/2	1/2	1	1/2	1	0	1/2	0	1/2	1/2	1	1/2	1
33	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
34	0	1/2	1/2	1	0	1/2	1/2	1	0	1/2	1/2	1	0	1/2	1/2	1
44	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

$\mathcal{H}$ =

	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
11	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇
12	◇	◇	◇	◇	□	□	□	□	■	■	■	■	◇	◇	◇	◇
13	◇	◇	□	□	◇	◇	□	□	■	■	◇	◇	■	■	◇	◇
14	◇	□	◇	□	◇	□	◇	□	■	◇	■	◇	■	◇	■	◇
22	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇
23	◇	◇	□	□	■	■	◇	◇	◇	◇	□	□	■	■	◇	◇
24	◇	□	◇	□	■	◇	■	◇	◇	□	◇	□	■	◇	■	◇
33	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇
34	◇	□	■	◇	◇	□	■	◇	◇	□	■	◇	◇	□	■	◇
44	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇

Figure 1: Gain matrix $\mathcal{G}$ and feedback matrix $\mathcal{H}$ for a 4-armed binary dueling bandits resulting in 10 non-duplicate actions and 16 possible outcomes.

$\mathcal{S}_{(12)}$ =

	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
□	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
◇	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1
■	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0

Figure 2: Signal matrix for action (12) for the same problem as in Figure 1.

3. Hierarchy and basic concepts of partial monitoring problems

In this section, firstly, we take a brief review of the basic concepts of partial monitoring problems. Most of the definitions in this section are taken from Bartók et al. (2011) and Bartók (2013).

Consider a finite partial monitoring game with action set $\boldsymbol{N}$, outcome set $\boldsymbol{M}$, loss matrix $\mathcal{L}$ and feedback matrix $\mathcal{H}$. For any action $i \in \boldsymbol{N}$, loss vector $\boldsymbol{l}_i$ denotes the column vector consisting of i^{th} row in $\mathcal{L}$. Correspondingly, gain vector $\boldsymbol{g}_i$ denotes the column vector consisting of i^{th} row in $\mathcal{G}$. For the rest of the article, gain vector $\boldsymbol{g}_i$ and loss vector $\boldsymbol{l}_i$ will be used interchangeably depending upon the setting. Let $\Delta_{|\boldsymbol{M}|}$ be the $|\boldsymbol{M}|-1$-dimensional probability simplex i.e. $\Delta_{|\boldsymbol{M}|} = \left\{\boldsymbol{q} \in [0,1]^{|\boldsymbol{M}|} \mid \|\boldsymbol{q}\|_1 = 1\right\}$. For any outcome sequence of length T, the vector $\boldsymbol{q}$ denoting the relative frequencies with which each outcome occurs is in $\Delta_{|\boldsymbol{M}|}$. The cumulative loss of action i for this outcome sequence can hence be described as follows:

$$\sum_{t=1}^{T} \mathcal{L}(i, J_t) = T \cdot \boldsymbol{l}_i^\top \boldsymbol{q}$$

The vectors denoting the outcome frequencies can be thought of as the opponent strategies. These opponent strategies determine which action is optimal i.e. the action with the lowest cumulative loss. This induces a *cell decomposition* on $\Delta_{|\boldsymbol{M}|}$.

Definition 1 (Cells) *The cell of an action i is defined as*

$$C_i = \left\{\boldsymbol{q} \in \Delta_{|\boldsymbol{M}|} \mid \boldsymbol{l}_i^\top \boldsymbol{q} = \min_{j \in \boldsymbol{M}} \boldsymbol{l}_j^\top \boldsymbol{q}\right\}$$

In other words, a cell of an action consists of those opponent strategies in the probability simplex for which it is the optimal action. An action i is said to be *Pareto-optimal* if there exists an opponent strategy $\boldsymbol{q}$ such that the action i is optimal under $\boldsymbol{q}$. The actions whose cells have a positive $(|\boldsymbol{M}|-1)$-dimensional volume are called *Strongly Pareto-optimal*. Actions that are Pareto-optimal but not strongly Pareto-optimal are called *degenerate*.

Definition 2 (Cell decomposition) *The cells of strongly Pareto-optimal actions form a finite cover of Δ_M called as the cell-decomposition.*

Two actions cells i and j from the cell decomposition are *neighbors* if their intersection is an $(|\boldsymbol{M}|-2)$-dimensional polytope. The actions corresponding to these cells are also called as *neighbors*. The raw feedback matrices can be 'standardized' by encoding their symbols in *signal matrices*:

Definition 3 (Signal matrices) *For an action i, let $\sigma_1, \ldots, \sigma_{s_i} \in \Sigma$ be the symbols occurring in row i of $\mathcal{H}$. The signal matrix $\mathcal{S}_i$ of action i is defined as the incidence matrix of symbols and outcomes i.e. $\mathcal{S}_i(k,m) = [\![\mathcal{H}(i,m) = \sigma_k]\!] \quad k = 1, \ldots, s_i, \quad$ for $m \in \boldsymbol{M}$* [4].

Observability is a key notion to assess the difficulty of a PM problem in terms of regret R_T against best action at time T.

4. we use $[\![\cdot]\!]$ to denote the indicator function

Definition 4 (Observability) *For actions i and j, we say that $\boldsymbol{l}_i - \boldsymbol{l}_j$ is* globally observable *if $\boldsymbol{l}_i - \boldsymbol{l}_j \in \operatorname{Im} \mathcal{S}^\top$. Where the global signal matrix $\mathcal{S}$ is obtained by stacking all signal matrices. Furthermore, if i and j are neighboring actions, then $\boldsymbol{l}_i - \boldsymbol{l}_j$ is called* locally observable *if $\boldsymbol{l}_i - \boldsymbol{l}_j \in \operatorname{Im} \mathcal{S}_{i,j}^\top$ where the local signal matrix $\mathcal{S}_{i,j}$ is obtained by stacking the signal matrices of all neighboring actions for i, j: $\mathcal{S}_k$ for $k \in \{k \in \boldsymbol{N} \mid C_i \cap C_j \subseteq C_k\}$.*

Theorem 1 (Classification of partial monitoring problems) *Let $(\boldsymbol{N}, \boldsymbol{M}, \boldsymbol{\Sigma}, \mathcal{L}, \mathcal{H})$ be a partial monitoring game. Let $\{C_1, \ldots, C_k\}$ be it's cell decomposition, with corresponding loss vectors $\boldsymbol{l}_1, \ldots, \boldsymbol{l}_k$. The game falls into the following four regret categories.*

- *$R_T = 0$ if there exists an action with $C_i = \Delta_{|\boldsymbol{M}|}$. This case is called* trivial.
- *$R_T \in \Theta(T)$ if there exist two strongly Pareto-optimal actions i and j such that $\boldsymbol{l}_i - \boldsymbol{l}_j$ is not globally observable. This case is called* hopeless.
- *$R_T \in \tilde{\Theta}(\sqrt{T})$ if it is not trivial and for all pairs of (strongly Pareto-optimal) neighboring actions i and j, $\boldsymbol{l}_i - \boldsymbol{l}_j$ is locally observable. This case is called* easy.
- *$R_T \in \Theta(T^{2/3})$ if $\mathscr{G}$ is not hopeless and there exists a pair of neighboring actions i and j such that $\boldsymbol{l}_i - \boldsymbol{l}_j$ is not locally observable. This case is called* hard.

4. Dueling bandits in the partial monitoring hierarchy

This section examines the place of the dueling bandit problem in the hierarchy of partial monitoring problems described above. Note that the existence of the REX3 algorithm (Gajane et al., 2015) with a $\tilde{\Theta}\left(\sqrt{KT}\right)$ regret guarantee is enough to state that dueling bandit is an *easy game* according to the hierarchy described in Theorem 1, but our aim here is to retrieve this result from the PM machinery.

Theorem 2 (Duelings bandits: locally observable) *In a binary utility-based dueling bandit problem with more than two arms, all the pairs of actions are locally observable.*

Proof Consider a dueling bandit problem as defined in Section 2 with binary gains and $K > 2$ arms. The signal matrix of any action $(i,j) \in \boldsymbol{N}$ is defined as follows:

$$S_{(i,j)}(\square, \boldsymbol{m}) = [\![\boldsymbol{m}_i < \boldsymbol{m}_j]\!], \quad S_{(i,j)}(\diamond, \boldsymbol{m}) = [\![\boldsymbol{m}_i = \boldsymbol{m}_j]\!], \quad S_{(i,j)}(\blacksquare, \boldsymbol{m}) = [\![\boldsymbol{m}_i > \boldsymbol{m}_j]\!]$$

In the following, we show that for any pair of actions (i,j) and (i',j'), $\boldsymbol{g}_{(i',j')} - \boldsymbol{g}_{(i',j')}$ is locally observable. For the sake of readability, let's consider $S^\blacksquare$, $S^\diamond$ and $S^\square$ to be the column vectors containing the rows pertaining to the symbols $\blacksquare$, $\diamond$ and $\square$ of the signal matrix S respectively. We consider the following two cases for the pair of actions which together cover all the possibilities:

- A pair of actions that share at-least one common arm:

1. Actions (i,k) and (k,j). For any binary gain outcome $\boldsymbol{m}$, we have :

$$\begin{aligned}\boldsymbol{g}_{(i,k)} - \boldsymbol{g}_{(k,j)} &= \left(\frac{\boldsymbol{m}_i + \boldsymbol{m}_k}{2} - \frac{\boldsymbol{m}_k + \boldsymbol{m}_j}{2}\right)_{\boldsymbol{m}\in\boldsymbol{M}} \\ &= 0.5\left([\![\boldsymbol{m}_i > \boldsymbol{m}_j]\!] - [\![\boldsymbol{m}_j > \boldsymbol{m}_i]\!]\right)_{\boldsymbol{m}\in\boldsymbol{M}} \\ &= 0.5\left(S^{\blacksquare}_{(i,j)} - S^{\square}_{(i,j)}\right) \qquad (1)\end{aligned}$$

So, $\boldsymbol{g}_{(i,k)} - \boldsymbol{g}_{(k,j)}$ falls in the row space of the signal matrix of the action (i,j) and hence in the row space of the signal matrix of the neighborhood action set. (refer definition 4)

2. Actions (i,k) and (j,k). Similarly, $\boldsymbol{g}_{(i,k)} - \boldsymbol{g}_{(j,k)} = 0.5S^{\blacksquare}_{(i,j)} - 0.5S^{\square}_{(i,j)}$.

- No common arm $(i \neq i' \neq j \neq j')$: In this case,

$$\begin{aligned}\boldsymbol{g}_{(i,j)} - \boldsymbol{g}_{(i',j')} &= \boldsymbol{g}_{(i,j)} - \boldsymbol{g}_{(i,j')} + \boldsymbol{g}_{(i,j')} - \boldsymbol{g}_{(i',j')} \\ &= 0.5\left(S^{\blacksquare}_{(j,j')} - S^{\square}_{(j,j')} + S^{\blacksquare}_{(i,i')} - S^{\square}_{(i,i')}\right) \qquad \text{Using equation (1)}\end{aligned}$$

Hence, for any pair of actions (i,j) and (i',j'), $\boldsymbol{g}_{(i,j)} - \boldsymbol{g}_{(i',j')}$ falls in the row space of the signal matrix of the neighborhood action set i.e. $\boldsymbol{g}_{(i,j)} - \boldsymbol{g}_{(i',j')} \in \operatorname{Im} S^{\top}_{((i,j)(i',j'))}$ and therefore it is locally observable. So, by extension, the binary dueling bandit problem is locally observable and hence we arrive at the following corollary. ■

Corollary 3 *According to the hierarchy described in theorem 1, the binary dueling bandit problem is* easy *and its regret is* $\tilde{\Theta}(\sqrt{T})$.

5. Partial monitoring algorithms and their use for dueling bandits

FEEDEXP3 by Piccolboni and Schindelhauer (2001) was the first algorithm for finite partial monitoring games. For its application, there is an important pre-condition – existence of a matrix $\mathcal{B}$ such that $\mathcal{B}\mathcal{H} = \mathcal{G}$. We prove by contradiction that such a matrix $\mathcal{B}$ doesn't exist for the dueling bandit problem. Let's assume $\mathcal{B}$ exists. Therefore, for any action $(i,j) \in \boldsymbol{N}$ and any outcome vector $\boldsymbol{m} \in \boldsymbol{M}$,

$$\mathcal{G}((i,j),\boldsymbol{m}) = \sum_{i',j'=1}^{K} \mathcal{B}_{((i,j)(i',j'))} \cdot \mathcal{H}_{((i',j')(\boldsymbol{m}))}$$

Consider $\boldsymbol{m} = 0\ldots0$, i.e. the gain of every arm is 0. In this case, the gain of any action (i,j) is 0 and the feedback for every action is $\diamond$, therefore

$$0 = \sum_{i',j'=1}^{K} \mathcal{B}_{((i,j)(i',j'))} \cdot \diamond \qquad (2)$$

Now consider $\boldsymbol{m} = 1\ldots1$, i.e. the gain of every arm is 1. In this case, the gain of any action (i,j) is 1 and feedback of every action is $\diamond$, therefore

$$1 = \sum_{i',j'=1}^{K} \mathcal{B}_{((i,j)(i',j'))} \cdot \diamond \qquad (3)$$

Eq. 2 and eq. 3 reach a contradiction, therefore our assumption that $\mathcal{B}$ exists is incorrect. Fortunately, the authors also provide a general algorithm which performs several matrices transformations to sidestep this pre-condition. These transformations are studied thoroughly in (Bartók, 2012).

BALATON by Bartók et al. (2011), CBP-vanilla and CBP by Bartók (2012) belong to the family of algorithms for the locally observable PM games as does GLOBAL-EXP3 by Bartók (2013). Although, for GLOBAL-EXP3, its regret bound of $\tilde{\mathcal{O}}(\sqrt{N'T})$ does not directly depend on the number of actions, but rather on the structure of games as N' is the size of the largest *point-local game*. We can however provide a counter-example for utility-based dueling bandits where $N' \approx K^2$ in the following way.

We use the notations from Bartók (2013). Consider a p in the probability simplex $\Delta_{|M|}$ where all the arms have maximal gains. For this p, all the actions are optimal therefore this point belongs to all the cells in the cell-decomposition. Hence, according to definition 6 in Bartók (2013), there exists a point-local game consisting of all the $K(K+1)/2$ non-duplicate actions. Therefore the upper bound of GLOBALEXP3 translates to $\tilde{O}(K\sqrt{T})$ for utility-based dueling bandits.

The following table summarizes the salient features of these PM algorithms. We can clearly see that none of them, except REX3, is optimal with respect to the number of actions N. Please note that for the dueling bandits problem, $N \approx K^2$.

Table 1: Summary of PM algorithms

Algorithm	**Setting**	**Optimality**	**Regret**
FEEDEXP3 (Piccolboni and Schindelhauer (2001))	Adversarial	Not in T or N	$\tilde{\mathcal{O}}(T^{2/3}K)$
BALATON (Bartók et al. (2011))	Stochastic	Not in T or N	$\tilde{\mathcal{O}}(K\sqrt{T})$
CBP (Bartók (2012))	Stochastic	in T, not in N	$\tilde{\mathcal{O}}(K^2 logT)$
GLOBAL-EXP3 (Bartók (2013))	Adversarial	in T, not in N	$\tilde{\mathcal{O}}(K\sqrt{T})$
SAVAGE (Urvoy et al. (2013))	Stochastic	in T, not in N	$\mathcal{O}(K^2 logT)$
NEIGHBORHOOD WATCH (Foster and Rakhlin (2011))	Adversarial	in T, not in N	$\tilde{\mathcal{O}}(K\sqrt{T})$
REX3 (Gajane et al. (2015))	Adversarial	in T and N	$\tilde{\mathcal{O}}(\sqrt{KT})$

6. Conclusion

In this article, we studied the dueling bandit problem as an instance of the partial monitoring problem. We proved that the binary dueling bandit problem is a locally observable game and hence falls in the easy category of the partial monitoring games. We also looked at the some of the existing partial monitoring algorithms and their optimality with respect to both time and the number of actions.

Table 2: Notation table

Notation	Description
K	Number of arms
t	Time index
T	Time horizon
R_T	Cumulative regret after time T
$\mathbb{E}_{\sim\pi}(\dots)$	Expectation according to π
$\boldsymbol{N}$	set of actions
$\boldsymbol{M}$	set of outcomes
$\boldsymbol{m}$	outcome vector $\in \boldsymbol{M}$
$\mathcal{L}$	loss function/matrix
$\mathcal{G}$	gain function/matrix
$\mathcal{H}$	feedback function/matrix
$\boldsymbol{l}_i$	loss vector: column vector consisting of i^{th} row in $\mathcal{L}$
$\boldsymbol{g}_i$	gain vector: column vector consisting of i^{th} row in $\mathcal{G}$
C_i	Cell of action i
$\lvert . \rvert$	size of set .
$\mathbb{R}$	Set of real numbers

References

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 263–274, 2008. URL `http://colt2008.cs.helsinki.fi/papers/127-Abernethy.pdf`.

Alekh Agarwal, Peter L. Bartlett, and Max Dama. Optimal allocation strategies for the dark pool problem. In Yee Whye Teh and D. Mike Titterington, editors, *AISTATS*, volume 9 of *JMLR Proceedings*, pages 9–16. JMLR.org, 2010. URL `http://dblp.uni-trier.de/db/journals/jmlr/jmlrp9.html#AgarwalBD10`.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Gábor Bartók. *The Role of Information in Online Learning*. PhD thesis, Edmonton, Alta., Canada, 2012. AAINR89904.

Gábor Bartók. A near-optimal algorithm for finite partial-monitoring games against adversarial opponents. In *Proc. COLT*, 2013.

Gábor Bartók, Dvid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *In Conference on Learning Theory*, 2011.

Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring - classification, regret bounds, and algorithms. *Math. Oper. Res.*, 39(4):967–

997, 2014. doi: 10.1287/moor.2014.0663. URL http://dx.doi.org/10.1287/moor.2014.0663.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. doi: 10.1561/2200000024. URL http://dx.doi.org/10.1561/2200000024.

Nicol Cesa-bianchi, Gbor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Trans. Inform. Theory*, 51:77–92, 2005.

Abraham Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *CoRR*, cs.LG/0408007, 2004. URL http://arxiv.org/abs/cs.LG/0408007.

Dean P. Foster and Alexander Rakhlin. No internal regret via neighborhood watch. *CoRR*, abs/1108.6088, 2011. URL http://arxiv.org/abs/1108.6088.

Pratik Gajane, Tanguy Urvoy, and Fabrice Clérot. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *ICML*, volume 37 of *JMLR: W&CP*, 2015.

Robert D. Kleinberg and Frank Thomson Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *FOCS*, pages 594–605. IEEE Computer Society, 2003. ISBN 0-7695-2040-5. URL http://dblp.uni-trier.de/db/conf/focs/focs2003.html#KleinbergL03.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261, February 1994. ISSN 0890-5401. doi: 10.1006/inco.1994.1009. URL http://dx.doi.org/10.1006/inco.1994.1009.

Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *COLT/EuroCOLT*, volume 2111 of *LNCS*, pages 208–223. Springer, 2001. ISBN 3-540-42343-5. doi: 10.1007/3-540-44581-1_14. URL http://dblp.uni-trier.de/db/conf/colt/colt2001.html#PiccolboniS01.

F. Radlinski and T. Joachims. Active exploration for learning rankings from clickthrough data. In *KDD 2007*, pages 570–579. ACM, 2007. doi: 10.1145/1281192.1281254.

Tanguy Urvoy, Fabrice Clerot, Raphael Fraud, and Sami Naamane. Generic exploration and k-armed voting bandits. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 91–99. JMLR Workshop and Conference Proceedings, May 2013. URL http://jmlr.csail.mit.edu/proceedings/papers/v28/urvoy13.pdf.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *ICML 2009*, pages 1201–1208. Omnipress, 2009. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553527.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 928–936, 2003. URL http://www.aaai.org/Library/ICML/2003/icml03-120.php.